

### Applied Statistics Qualifier Exam

This exam consists of 2 parts. You must answer 5 questions total and must answer at least 2 questions from each part. Make sure to clearly indicate which problems you are attempting.

#### PART A

A1. Consider a random effect model

$$E(Y_{ij}|a_i) = \mu + a_i$$

where  $Y_{ij}|a_i$ s are independent and each follows Normal distribution with mean  $\mu + a_i$  and variance  $\sigma^2$ . Also,  $a_i$ s are independent and each follows Normal distribution with mean 0 and variance  $\sigma_a^2$ . Calculate  $Cov(Y_{ij}, Y_{ik})$  when  $j \neq k$ . Also, calculate  $Var(Y_{ij})$ .

A2. Consider a Beta-Binomial model

$$E(Y_{ij}|p_i) = p_i$$

where  $Y_{ij}|p_i$ s are independent and each follows Bernoulli distribution with success probability  $p_i$ . Also,  $p_i$ s are independent and each follows Beta distribution with parameters  $(\alpha, \beta)$ . Calculate  $\rho = Corr(Y_{ij}, Y_{ik})$  when  $j \neq k$ .

A3. Show that  $E[Y|X]$  is the minimum mean square error predictor of  $Y$ . That is, show that  $g(X) = E[Y|X]$  minimizes  $E[(Y - g(X))^2]$  among all functions  $g(\cdot)$  of  $X$ .

A4. Show that

$$\frac{1}{6} \begin{bmatrix} 1 & 16 & 9 & -6 \\ -1 & -14 & -9 & 6 \\ -1 & -16 & -6 & 6 \\ -1 & -16 & -9 & 12 \end{bmatrix}$$

is a generalized inverse of  $\mathbf{X}^T \mathbf{X}$  where

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

PART B

B1. Consider the model

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

for  $i = 1, \dots, n$  where the  $\epsilon_i$ 's are independent and identically distributed normal random variables with mean 0 and unknown constant variance  $\sigma^2 > 0$ . Now suppose we observe the following data:

$x$	-2	-1	0	1
$y$	6	1	1	5

- (a) Find the maximum likelihood estimator of  $\boldsymbol{\beta} = (\alpha, \beta_1, \beta_2)'$ .
- (b) Test the hypothesis that  $H_0 : \beta_2 = 0$  at a 5% level of significance using the likelihood ratio test. The tables below give the 95th and 97.5th percentile of the  $F$  distribution with  $df1$  and  $df2$  degrees of freedom. (Hint: For testing  $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ , the likelihood-based  $F$ -statistic has the forms

$$\begin{aligned}
 F &= \frac{(\text{reduced SS} - \text{full SS})/q}{\text{full SS}/(N - p)} \\
 &= \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)\|^2/q}{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(N - p)} \\
 &= \frac{(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1}(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m})/q}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N - p)} \sim f_{q, N-p}
 \end{aligned}$$

where the (restricted) MLE of  $\boldsymbol{\beta}$  is denoted as  $\hat{\boldsymbol{\beta}}_0$ , reduced SS =  $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0\|^2$  and full SS =  $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$ .

		95%			
		$df1$			
		1	2	3	4
$df2$	1	161.447	199.500	215.707	224.583
	2	18.513	19.000	19.164	19.247
	3	10.128	9.552	9.277	9.117
	4	7.709	6.944	6.591	6.388

		97.5%			
		$df1$			
		1	2	3	4
$df2$	1	647.789	799.500	864.163	899.583
	2	38.506	39.000	39.165	39.248
	3	17.443	16.044	15.439	15.101
	4	12.218	10.649	9.979	9.605

B2. Suppose we have an input variable  $X$  that we wish to use for classification. The output variable  $Y$  is a Bernoulli random variable that follows the no-intercept logistic regression model

$$\ln \frac{\Pr(Y = 1|X = x)}{\Pr(Y = 0|X = x)} = \beta x.$$

Given training data  $(x_1, y_1), \dots, (x_N, y_N)$  where  $y_1, \dots, y_N \in \{0, 1\}$ , the log-likelihood function for the  $N$  observations is given by

$$\ell(\beta) = \sum_{i=1}^N \ln \Pr(Y = y_i | X = x_i).$$

(a) Show that

$$\frac{d\ell}{d\beta} = \sum_{i=1}^N x_i \left( y_i - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right).$$

Now suppose we observe the following 3 observations:  $(-1, 1), (0, 0), (1, 1)$ .

(b) Find the maximum likelihood estimate of  $\beta$ .

(c) Show that the solution of (b) is the unique maximizer of  $\ell$ .

B3. Suppose we observe  $N$  data points  $(x_i, y_i), i = 1, \dots, N$ , and we want to minimize the penalized residual sum of squares

$$V(\beta) = \sum_{i=1}^N (y_i - \beta x_i)^2 + \lambda \beta^2.$$

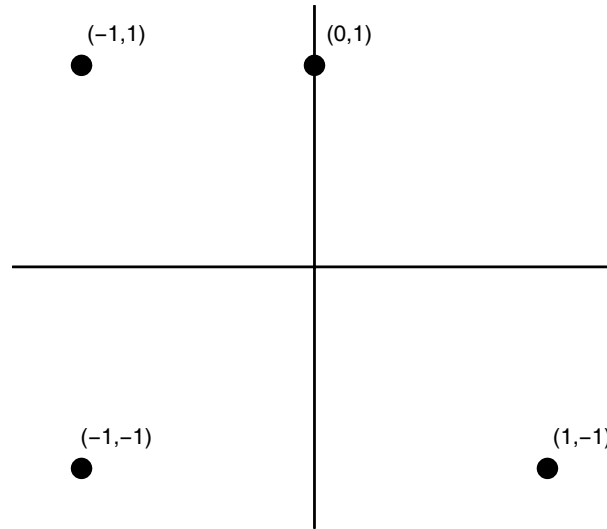
where  $\beta$  is an unknown constant and  $\lambda > 0$  is a penalty parameter under this no intercept model.

(a) If  $\lambda$  is known, what value of  $\beta$  minimizes  $V$ ? Justify that your solution is a minimizer.

(b) Denote the estimator of  $\beta$  obtained in part (a) as  $\hat{\beta}(\lambda)$ . Under the no-intercept regression model  $Y_i = \beta x_i + \epsilon_i$  for  $i = 1, \dots, N$  where the  $x_i$ 's are known non-random inputs, and the  $\epsilon_i$ 's are independent and identically distributed random variables with mean 0 and constant variance  $\sigma^2 > 0$ ,

1. what is the bias and variance of  $\hat{\beta}(\lambda)$ ?
2. find the value of  $\lambda$  for which  $\hat{\beta}(\lambda)$  has the smallest mean squared error  $E[(\hat{\beta}(\lambda) - \beta)^2]$ ? Does this depend on  $\beta$ ?  $\sigma^2$ ?

B4. Consider the data shown below:



- (a) Apply the  $K$ -means algorithm to obtain 2 clusters by beginning with cluster A centered at  $(-1, 1)$  and cluster B centered at  $(0, 1)$ .
- (b) Apply the  $K$ -means algorithm to obtain 2 clusters by beginning with cluster A centered at  $(-1, 1)$  and cluster B centered at  $(-1, -1)$ .
- (c) Apply the  $K$ -means algorithm to obtain 2 clusters by beginning with cluster A centered at  $(0, 0)$  and cluster B centered at  $(1, -1)$ .
- (d) Which initial arrangement gives the best result?